Initial Validation of a Test of Spatial Knowledge in Anatomy

Charles P. Friedman[1], Parvati Dev[2], Bonnie Dafoe[2], Gwendolyn Murphy[1], Ramon Felciano[2]
[1]University of North Carolina School of Medicine, Chapel Hill, NC
[2]Stanford University School of Medicine, Stanford, CA

Abstract: The authors have developed HERCULES, a computer-based test designed to assess the spatial, non-verbal components of knowledge in anatomy. The test consists of two tasks, each requiring subjects to estimate the vertical level in the body of a set of color, cross-sectional images. In Task 1, subjects make the estimate based on a limited number of clues, where each clue is an anatomical structure that appears in the cross-section. In Task 2, subjects estimate the level based on a view of the cross-section with all structures shown. A validation study of this test using six images for each task was performed with preclinical medical students, fourth year medical students, and experienced teachers of anatomy as subjects. Results indicate that the exercise is at an appropriate level of difficulty and that a somewhat longer test than used in this study would be adequately reliable for use in actual assessment. The test appears to discriminate the expert faculty from more novice students and thus exhibits an aspect of validity that is very important in assessment exercises of this type.

BACKGROUND

Recent advances in the ability of microcomputers to display and manipulate high-resolution color images have spawned a new generation of computer-based educational programs in anatomy [1-4]. These programs allow students to navigate visually through regions of the body, to display structures at differing levels of magnification, and to perform electronic dissection by, for example, "cutting through" an organ and then rotating the resulting two pieces to observe them in cross-section. Because these programs are so visually powerful, they might be expected to develop student's spatial knowledge of anatomy.

In addition to its tutorial/instructional uses, this new technology can also be applied to test the kinds of anatomical knowledge students might be expected to derive through their interactions with these programs. Such tests are necessary to assess students appropriately if these programs are used in routine instruction, and they would be an essential component of studies to determine the educational value of these programs. This paper reports an empirical validation of HERCULES, a computer-based test explicitly designed to assess spatial knowledge of anatomy [5].

Knowledge of anatomy is known to have both verbal and spatial/visual components [6]. Purely verbal knowledge, for example, might include the classifications of specific structures. Other knowledge has both verbal and spatial components; for example, the ability to recognize and name a structure whose image is displayed. Purely spatial knowledge involves the locations and orientations of multiple objects in three dimensions: an understanding of how the body is put together. Spatial knowledge is known to be encoded and retrieved through cognitive processes different from those applying to verbal knowledge [7,8]. Proficiency in one area does not imply proficiency in the other. Building on Rochford's example [6], a student who knows that the corpus callosum can be classified as a commissure may not be able to place the corpus callosum within a given cross-section or diagram of the brain--and may not be able to determine whether the corpus callosum, if not explicitly shown, belongs in that cross-section. Several studies [6,9] have linked substandard medical student performance in anatomy to deficits in visual memory and general psychological measures of spatial reasoning.

HERCULES tests spatial knowledge of anatomy using cross-sectional images taken from Peterson [10]. The general task is to determine the correct vertical level of a cross-section of the thorax, abdomen, or pelvis. Two versions of this general task are the foci of this study. In Task 1, subjects are challenged to determine the vertical level of a cross-section based on an incomplete display of the structures found in that cross-section. Subjects are initially shown only the border of the cross-section. One anatomical structure (e.g. lung, vertebra, muscle) appears in color and in its proper location within the cross-section each time the subject requests a "clue." After each new clue is displayed, subjects give their best estimate of the level of the cross-section by clicking at the desired level of a diagrammatic person that is always visible in frontal view. The first three clues available for each cross-sectional image contain what the investigators believe to be relatively little spatially-orienting information; they are typically muscle and fatty tissue. Later clues, which appear in a randomized order, contain structures, such as major organs, that are more informative. Subjects are told that their score for each image is based on the accuracy of their final estimate as well as the number of clues (structures) on which their final estimate is based. The program displays the value of an "attenuation factor" that indicates to subjects how much the next clue will reduce their maximum possible score and in this way subjects are discouraged from taking further clues once they are reasonably certain of the level of the cross-section.

In Task 2, subjects are shown a set of complete cross-sectional images--each with all structures displayed--and asked to estimate each image's vertical level. In the second task, the subject's score is based only on the difference between the subject's estimated level and the true level of the image.

HERCULES was developed by the investigators in SuperCard to run on Macintosh II series computers with color video display. Technical aspects of the HERCULES program are described elsewhere [5].

Before HERCULES can be appropriately used in educational settings, its extent of reliability and validity must be determined. This is especially necessary for an exercise such as HERCULES which is measuring a non-verbal ability using a computer-based format with which there has been little previous experience. A particularly important aspect of validity for assessments of this type compares the performance of novices and experts. Other prototype assessment tools in medical education have failed this test [11].

RESEARCH QUESTIONS

This study explores several psychometric properties of HERCULES, specifically:

1) Are students and faculty able to complete the tasks, and do their scores fall into ranges which suggest that the tasks are at an appropriate level of difficulty?

2) How reliable are the measures provided by each HERCULES task?

3) Are the measures provided by HERCULES valid in the sense that performance increases with increased experience in anatomy? Which task better discriminates the experts and novices?

4) To what extent are the two HERCULES tasks measuring different attributes?

We specifically hypothesized that Task 1, which requires estimation based on incomplete visual information, will better discriminate novice students and expert anatomists than Task 2 which is based on complete cross-sections.

METHOD

Subjects for the study were medical students and faculty members from Stanford University and the University of North Carolina: 10 preclinical medical students who had completed their gross anatomy course, 13 fourth year medical students, and 10 active doctoral-level teachers of gross anatomy. All subjects were volunteers who responded to a general solicitation. As such, they comprised a convenience sample appropriate to an initial validation study. After an introduction to the program by a research assistant, all subjects completed both HERCULES tasks in one sitting. Each task comprised six cross-sectional images. The research assistant was present for the entire period of work, to assist subjects who might encounter difficulties with the program and to conduct a brief interview of each subject after s/he completed the program.

The program automatically generated a file containing a complete record of each subject's performance on HERCULES. For the first (incremental clues) task, the investigators created a scoring algorithm suited to the structure of the exercise. Each subject's score for each image was based on the accuracy of the subject's final estimate multiplied by an attenuation factor inversely proportional to the number of clues on which the final estimate was based:

$$\text{Score} = (\text{Attenuation Factor}) \times (10 - |\text{Correct Level} - \text{Final Estimate of Level}|)$$

The attenuation factor is unity for four or fewer clues and then decreases linearly, as a function of the total number of clues available in the image. Thus, the maximum score for each image was 10, representing a perfectly accurate guess based on four or fewer clues. For an image containing 11 total clues, a

	Preclinical Students	Fourth Year Students	All Students	Faculty
Task 1				
Mean	47.2	50.5	49.0	62.7
SD	11.2	15.3	13.3	5.2
N	10	12	22	8
Task 2				
Mean	76.0	69.8	72.5	80.2
SD	10.0	14.5	12.8	5.8
N	10	13	23	10

Table 1: Performance on HERCULES Tasks as Function of Experience Level
(Note: All scores are reported as percentages.)

subject whose final estimate was incorrect by three levels based on six clues would receive a score for the image of .87 x 7 = 6.09. Final estimates with errors of greater than nine levels received a score of zero. For the second task, all clues are shown simultaneously, so the scoring is as above but with the attenuation factor always equal to one. For each task, each subject's total score was the sum of the scores on each of the six images.

The reliability of each HERCULES task was computed by considering each image as one item of a conventional test, so each task had six "items." Cronbach's alpha was computed as an index of reliability. Validity was explored using analysis of variance to compare the mean scores for each group: faculty, fourth year students, and preclinical students. A secondary analysis compared the means for faculty with the means for all students combined. The overlap of the traits measured by the two tasks was estimated by Pearson correlation.

RESULTS

All subjects completed the tasks without apparent difficulty. Preliminary analysis of the results revealed that two faculty members and one fourth year student appeared to misunderstand the intent of the first task since these individuals never selected more than four clues for any image. Since these individuals completed the exercise using a strategy different from what was intended, and in a manner not comparable with the other subjects, we deleted their Task 1 results from subsequent analyses.

The scores of the subjects overall and for each level of experience are reported in Table 1. The mean scores ranged from 80.2% for the faculty group on the second task, to 47.2% for the preclinical students on the first task. Student scores display considerable variability and more variability than faculty scores. All scores appear to fall into a useful range for assessment.

The reliabilities of the two tasks were .413 and .563 respectively. Combining the two tasks into a single "battery" comprised of 12 images increased the reliability to .677.

With reference to Table 1 and for Task 1, the differences between the means for each group are statistically significant ($F(2,27)=4.014$, $p<.05$), with faculty displaying the highest scores followed by the fourth year students and preclinical students. Faculty scores were significantly higher than those for all students combined ($t=4.023$, $df=27.7$, $p<.001$). For Task 2, the global differences between the three groups were not significant ($F(2,30)=2.468$, $p=.10$) although the faculty scored significantly higher than the students as a group ($t=2.332$, $df=30.8$, $p<.05$).

The correlation between Task 1 and Task 2 scores was .665 ($p<.01$).

DISCUSSION

This initial validation study suggests that HERCULES is of potential value as an assessment tool in anatomy. Mean scores for students on Task 1 are close to 50%, which is optimal for discriminating individual ability levels. Task 2 is easier for both students and faculty, but still allows discrimination. Although the reliabilities reported in this study are modest, they can be increased by lengthening the test.

For a test with 12 images per task, instead of the six used in this study, the Spearman-Brown prophecy formula yields a predicted reliability of .81.

The significant differences between student and faculty scores begin to make a case for the validity of the exercise as an assessment of anatomical knowledge. Task 1, which requires identification of anatomical levels based on parsimonious information, differentiates students and faculty to a greater degree than Task 2, where subjects work exclusively with complete images. The trend for preclinical students to outscore fourth year students on Task 2, but not on Task 1, is interesting and worthy of further exploration. As discussed earlier, this form of construct validity cannot be assumed in the assessment of medical students and professionals. Branching patient management problems, which displayed a substantial degree of face validity, have failed to distinguish medical trainees from experienced physicians [11].

Tasks 1 and 2 appear to be measuring somewhat different traits, but the "true" correlation between the scores is likely higher than the observed correlation of .665 due to the unreliability of both measures. A more definitive exploration of this relationship will require larger samples of subjects and a longer test.

Of course, the results reported here reflect to some degree the specific scoring system adopted by the authors, especially since the scoring system can directly affect subjects' behavior on Task 1. In general, subjects appeared to be strongly influenced by the score attenuation factor, the instantaneous value of which was displayed to them while working through the program. This feedback appeared to make them "clue averse," causing them in some cases to cease work on an image prematurely, making very inaccurate final estimates that would likely have been improved had they opted for one or two more clues. Subjects also expressed frustration when, after "paying" for a clue with a decrease in their maximum possible score, the resulting clue yielded what for them was little additional information. Further work is necessary to quantify the effects of the scoring algorithm on performance in these exercises, and to determine an optimal scoring system.

We plan to continue validation studies of the existing two HERCULES tasks, and also to develop at least one new task. In this new task, subjects will be presented initially with the outline of the cross-section. When they click somewhere in the cross-section, the program will display as a clue whichever structure exists at that location. As in Task 1 described above, subjects will be asked to make their best estimates based on a minimum of information, but they will have more control over what information they receive. While all of the HERCULES tasks to date assess spatial knowledge based on cross-sectional anatomy, we plan also to experiment with other formats for displaying anatomical structures in tests of spatial knowledge. These may have different psychometric properties when made into assessment exercises.

References

1. Eno K, Sundsten JW, Brinkley JF. A Multimedia Anatomy Browser Incorporating a Knowledge Base and 3D Images. *Symposium on Computer Applications in Medical Care,* **15**: 727-731, 1991.

2. Chapman CM, Miller JG, Bush LC, Bruenger JA, Wysor WJ, Meininger ET, Wolf FM, Fisher TV, Beaudoin AR, Burkel WE, MacCallum DK, Fisher DL, Carlson BM. ATLAS-*plus:* Multimedia Instruction in Embryology, Gross Anatomy, and Histology. *Symposium on Computer Applications in Medical Care,* **16**: 712-716, 1992.

3. McCracken TO, Spurgeon TL. The Vesalius Project: Interactive Computers in Anatomical Instruction. *Journal of Biocommunication,* **18** N2: 40-44, 1991.

4. Guy JF, Frisby AJ. Using Interactive Videodiscs to Teach Gross Anatomy to Undergraduates at the Ohio State University. *Academic Medicine,* **67**: 132-133, 1992.

5. Dev P, Friedman C, Dafoe B, Felciano R. Testing Spatial Understanding of Anatomy. *Symposium on Computer Applications in Medical Care,* **16**: 804-805, 1992.

6. Rochford K. Spatial Learning Disabilities and Underachievement among University Anatomy Students. *Medical Education,* **19**: 13-26, 1985.

7. Treisman A. Features and Objects in Visual Processing. *Scientific American,* **225**: 114B-125, 1986.

8. Juhel J. Spatial Abilities and Individual Differences in Visual Information Processing. *Intelligence*, **15**: 117-137, 1991.

9. Folan JC, de Montfort Supple M. Visual Memory and Auditory Recall in Anatomy Students. *Medical Education*, **20**: 516-520, 1986.

10. Peterson RR. A Cross-sectional Approach to Anatomy. Published by R. R. Peterson, 1982.

11. Newble DI, Moore J, Baxter A. Patient Management Problems: Issues of Validity. *Medical Education*, 16: 137-142, 1982.